3D Reconstruction using Hybrid Depth and Tracking

Sai Arunkumar, Hitang Ketankumar Desai, Muhammad-Abdullah Kashif-Khan, Jaden Tellis, and Laleendteshwaran Vigneshwaran | Supervisor - Sharib Ali

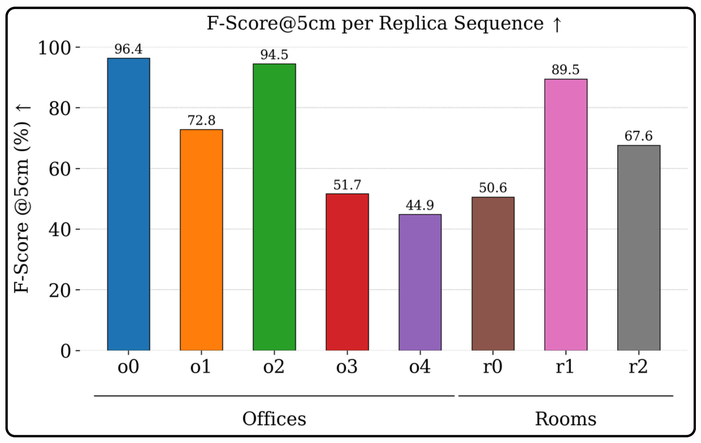
UNIVERSITY OF LEEDS

INTRODUCTION

Reconstructing 3D scenes typically requires depth sensors, calibrated cameras, or dedicated capture setups. We show that a single video from a consumer phone is enough. Our pipeline combines a frozen monocular depth model and a long-range point tracker with classical geometric optimisation, producing metric-scale textured meshes without any per-scene training.

No LiDAR. | No rig. | No second camera. | **Just One Video**

RESULTS



On the Replica dataset [4], beyond the reported F@5cm result, our pipeline reaches ***F@10cm of 80.8%*** and ***F@15cm of 88.3%***, indicating stronger surface agreement at broader tolerances. It also achieves an average ***ATE RMSE of 1.56cm***, showing stable trajectory alignment, with a mean Chamfer distance of 2.40cm. The strongest reconstructions occur on dominant room-scale surfaces, with future refinement focused on improving fine structures, dense clutter and partially observed regions.

71.0% MEAN F-SCORE @5cm

Average across 8 Replica sequences. Reconstructed surfaces lie within 5 cm of ground truth, at 1200×680 resolution

SINGLE-PASS RECONSTRUCTION

Video in, metric mesh + trajectory out. No manual seeding, no intermediate user input, no post-processing.

~16 mins PER SEQUENCE

End-to-end reconstruction of a 2,000-frame video on a single GPU. No real-time inference, no cluster required.

PIPELINE

1. INPUT: MONOCULAR RGB VIDEO



DAv3

2A. DEPTH ESTIMATION



- Predicts per-pixel metric depth
- Provides dense geometric structure

CoTracker3

2B. FEATURE TRACKING



- Tracks points across the entire video
- Ensures spatial correspondences

3. IRLS Filtering

- Iterative Reweighted Least Squares
- Removes outlier tracks
- Down-weights noisy observations

HUBER LOSS

$$H_\delta(e) = \begin{cases} \frac{1}{2}e^2, & |e| \le \delta \\ \delta\left(|e| - \frac{1}{2}\delta\right), & |e| > \delta \end{cases}$$

- Limits outlier track impact

4. GLOBAL BUNDLE ADJUSTMENT

- Optimises camera poses and depth maps
- Minimises reprojection error

5. TSDF FUSION

- Integrates depth over time
- Produces consistent volumetric map

6. OUTPUT: FINAL 3D RECONSTRUCTION



*Picture taken from Tanks and Temples Dataset [2]

- 3D Mesh & Point Cloud
- Accurate Geometry
- Reduced Drift
- Refined Trajectory

FUTURE APPLICATIONS

Heritage Capture | Colonoscopy & Surgical AR | Robotics | Forensics & Scene Documentation | Game Asset Capture | Drone-Based Site Surveying

The current system operates without additional training, while future fine-tuning of DAv3 or CoTracker3 could further improve performance in specialised video-based reconstruction tasks.

PERCEPTION MODULES

Depth Anything V3 (DAv3)

State-of-the-art monocular depth model pretrained on ~ 62M images. DAv3 predicts dense metric depth, confidence, and rough pose from RGB frames, providing strong frozen priors that later geometry stages align across the video [3].

CoTracker3

Transformer-based long-range point tracker. CoTracker3 follows thousands of 2D points through occlusion, motion, and viewpoint change, producing persistent correspondences that allow the geometry stages to link the same scene points across frames [1].

GROUND TRUTH VS RECONSTRUCTION

Our pipeline recovers major scene layout, object placement, and dominant structures across different indoor Replica environments

Ground Truth - Office



Reconstruction - Office



Ground Truth - Room



Reconstruction - Room



OPTIMISATION

The perception modules provide useful but imperfect signals: dense monocular depth gives per-frame geometry, while long-term tracks connect the same scene points across frames. Our pipeline brings these together using multi-view reprojection consistency, checking whether tracked points can be explained by one rigid 3D scene. Rather than retraining the models, it refines only geometric variables: camera poses, global scale and per-frame depth corrections. A fast rigidity filter removes inconsistent tracks, while global bundle adjustment refines camera trajectories and depth alignment by minimising reprojection error, producing corrected depth maps and a consistent trajectory for TSDF fusion.

REFERENCES

1. Karaev, N., Makarov, I., Wang, J., Neverova, N., Vedaldi, A. and Rupprecht, C. 2024. [Pre-print]. CoTracker3: Simpler and better point tracking by pseudo-labelling real videos. arXiv. [Online]. Available from: https://arxiv.org/abs/2410.11831. [Accessed 8 May 2026].
2. Knapitsch, A., Park, J., Zhou, Q.Y. and Koltun, V. 2017. Tanks and temples: Benchmarking large-scale scene reconstruction. ACM Transactions on Graphics. [Online]. 36(4), article no: 78. Available from: https://doi.org/10.1145/3072959.3073599. [Accessed 8 May 2026].
3. Lin, H., Chen, S., Liew, J.H., Chen, D.Y., Li, Z., Shi, G., Feng, J. and Kang, B. 2025. [Pre-print]. Depth Anything 3: Recovering the visual space from any views. arXiv. [Online]. Available from: https://arxiv.org/abs/2511.10647. [Accessed 8 May 2026].
4. Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J.J., Mur-Artal, R., Ren, C., Verma, S., Clarkson, A., Yan, M., Budge, B., Yan, Y., Pan, X., Yon, J., Zou, Y., Leon, K., Carter, N., Briales, J., Gillingham, T., Mueggler, E., Pesqueira, L., Savva, M., Batra, D., Strasdat, H.M., De Nardi, R., Goesele, M., Lovegrove, S. and Newcombe, R.A. 2019. [Pre-print]. The Replica dataset: A digital replica of indoor spaces. arXiv. [Online]. Available from: https://arxiv.org/abs/1906.05797. [Accessed 8 May 2026].